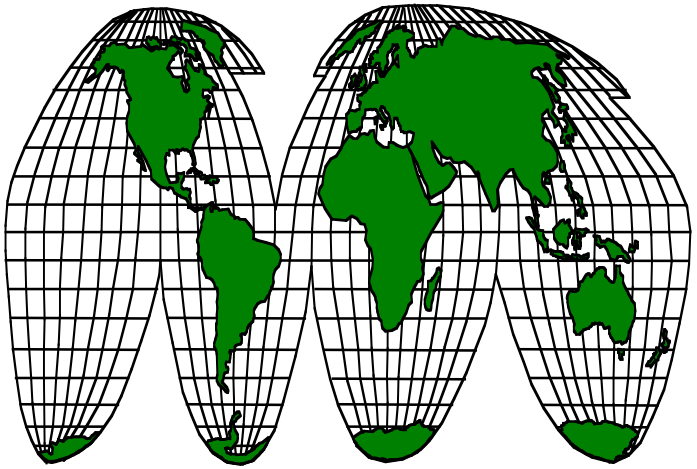
MONARC Project Status Report

http://www.cern.ch/MONARC



Harvey Newman
California Institute of Technlogy

http://l3www.cern.ch/monarc/monarc_lehman151100.ppt

Fromthe Talk at the DOE/NSF Review at BNL

November 15, 2000

MONARC History

Spring 1998	First Distributed Center Models (Bunn; Von Praun)
6/1998	Presentation to LCB; Project Assignment Plan
Summer 1998	MONARC Project Startup (ATLAS, CMS, LHCb)
9 - 10/1998	Project Execution Plan; Approved by LCB
1/1999	First Analysis Process to be Modeled
2/1999	First Java Based Simulation Models (I. Legrand)
Spring 1999	Java2 Based Simulations; GUI
4/99; 8/99; 12/99	Regional Centre Representative Meetings
6/1999	Mid-Project Progress Report Including MONARC Baseline Models
9/1999	Validation of MONARC Simulation on Testbeds Reports at LCB Workshop (HN, I. Legrand)
1/2000	Phase 3 Letter of Intent (4 LHC Experiments)
2/2000	Six Papers and Presentations at CHEP2000: *D385, F148, D127, D235, C113, C169*
3/2000	Phase 2 Report
Spring 2000	New Tools: SNMP-based Monitoring; S.O.N.N.
5/2000	Phase 3 Simulation of ORCA4 Production; Begin Studies with Tapes
Spring 2000	MONARC Model Recognized by Hoffmann WWC Panel; Basis of Data Grid Efforts in US and Europe


MONARC Working Groups/Chairs

Analysis Process Design WG — **P. Capiluppi (Bologna, CMS)**

Studied the analysis workload, job mix and profiles, time to complete the reco. and analysis jobs. Worked with the Simulation WG to verify that the specified resources in the models could handle the workload.

Architectures WG — **Joel Butler (FNAL, CMS)**

Studied the site and network architectures, operational modes and services provided by Regional Centres, data volumes stored and analyzed, candidate architectures for CERN, Tier1 (and Tier2) Centres

Simulation WG — **K. Sliwa (Tufts, ATLAS)**

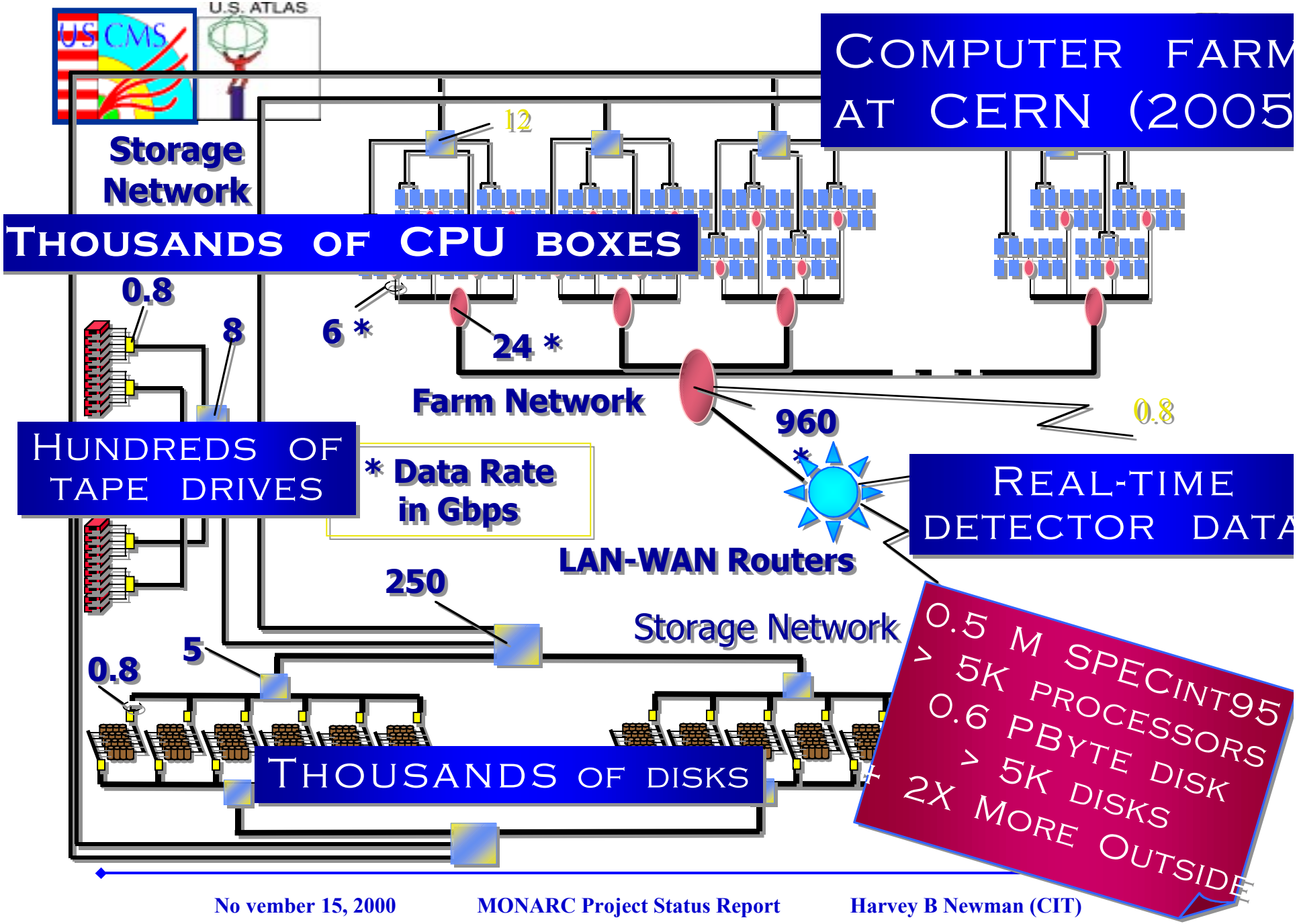
Defined the methodology, then (I. Legrand et al.) designed, built and further developed the simulation system as a toolset for users. Validated the simulation with the Testbeds group.

Testbeds WG — **L. Luminari (Rome, ATLAS)**

Set up small and larger prototype systems at CERN, several INFN and US sites and Japan, and used them to characterize the performance of the main elements that could limit throughput in the simulated systems

Steering Group — **Laura Perini (Milan, ATLAS)**, **Harvey Newman (Caltech, CMS)**

➡ **Regional Centres Committee**


U.S. ATLAS
US CMS
Storage Network
Computer farm at CERN (2005
Thousands of CPU boxes
12
0.8
8
6 *
24 *
Farm Network
960 *
0.8
Hundreds of tape drives
* Data Rate in Gbps
LAN-WAN Routers
Real-time detector data
250
Storage Network
5
0.8
Thousands of disks
0.5 M SPECINT95
> 5K processors
0.6 PByte disk
> 5K disks
2X More Outside

MONARC Analysis Model

Hierarchy of Processes (Experiment, Analysis Groups, Individuals)

Experiment-Wide Activity (10^9 events)

Reconstruction

3000 SI95sec/event
1 job year

Re-processing
3 per year

3000 SI95sec/event
3 jobs per year

New detector calibrations
Or understanding

Monte Carlo

5000 SI95sec/event

25 SI95sec/event
~20 jobs per month

~20 Groups' Activity (10^9 ➡ 10^7 events)

Selection

Iterative selection
Once per month

Trigger based and Physics based refinements

10 SI95sec/event
~500 jobs per day

~25 Individual per Group Activity (10^6 –10^8 events)

Analysis

Different Physics cuts & MC comparison
~1 time per day

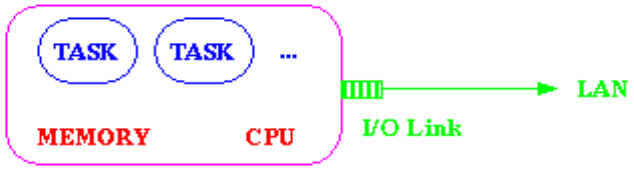
Algorithms applied to data to get results

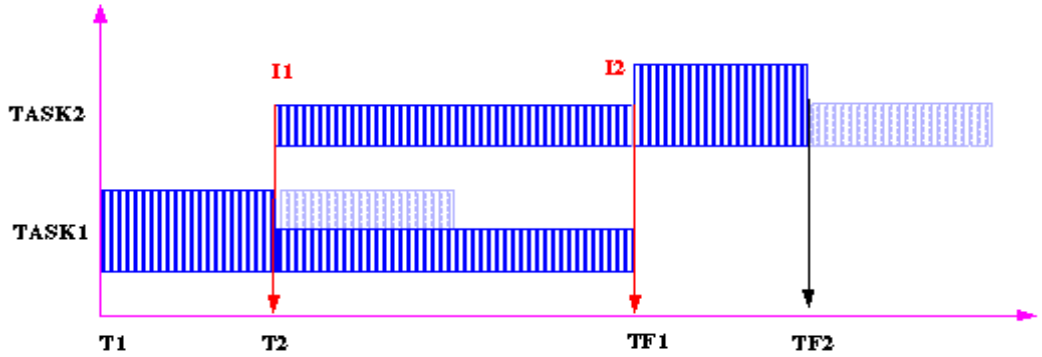
MONARC Simulation System Multitasking Processing Model

- Assign active tasks (CPU, I/O, network) to Java threads
- Concurrent running tasks share resources (CPU, memory, I/O)

"Interrupt" driven scheme:

For each new task or when one task is finished, an interrupt is generated and all "times to completion" are recomputed.



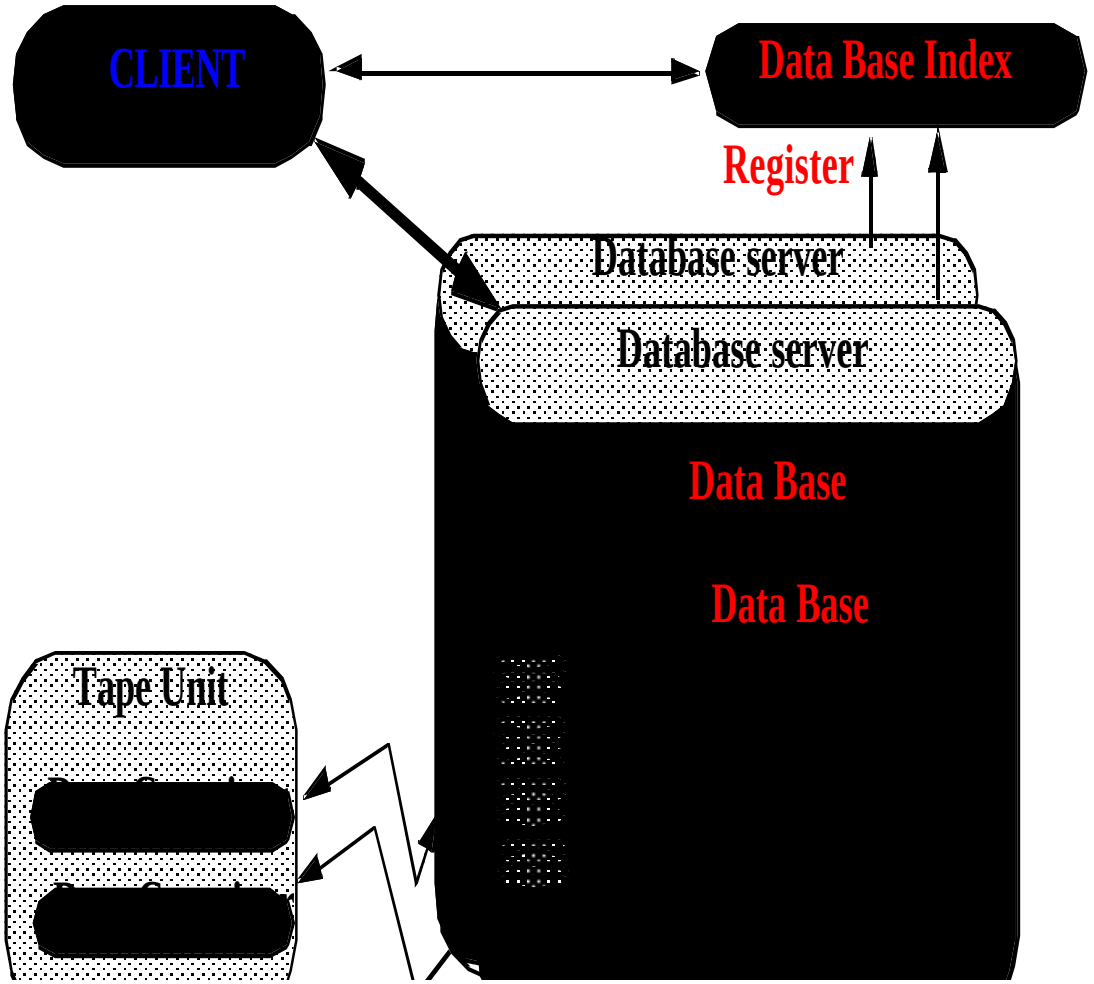


It provides:

An efficient mechanism to simulate multitask processing

An easy way to apply different load balancing schemes

MONARC Simulation Data Model

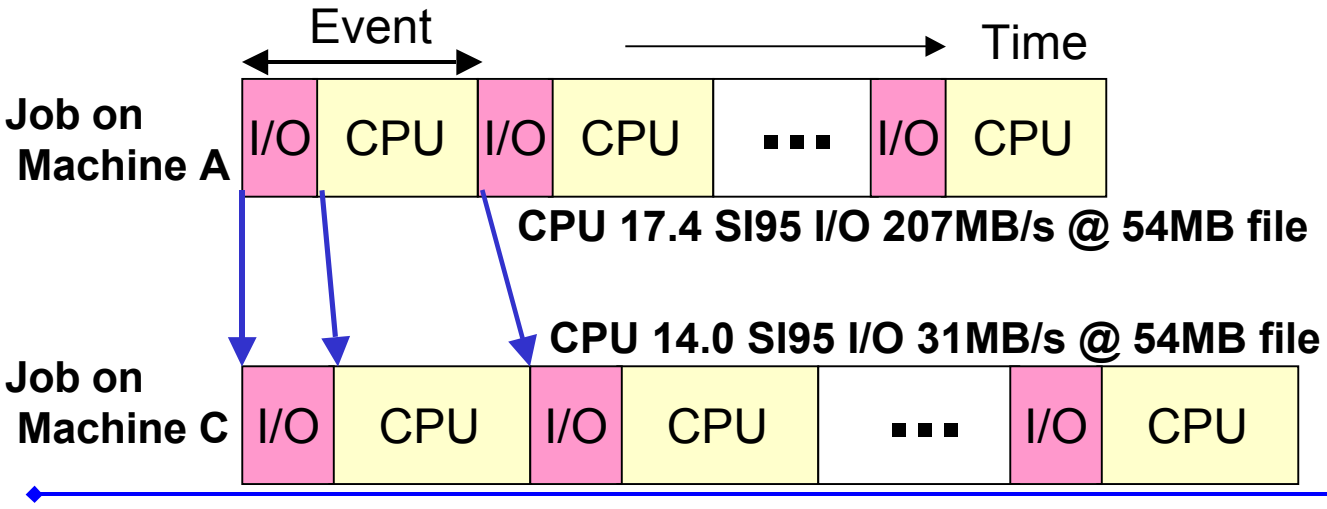


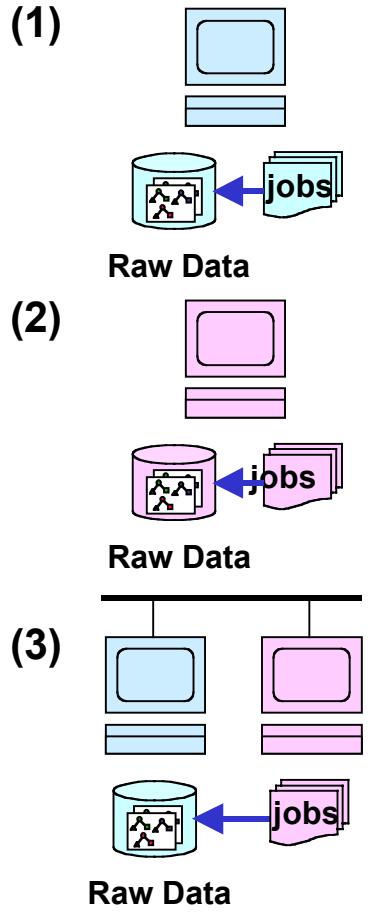
It provides

- Realistic mapping for an object data base
- Specific HEP data structure
- Transparent access to any data
- Automatic storage management
- An efficient way to handle very large number of objects.
- Emulation of clustering factors for different types of access patterns.
- Handling related objects in different data bases.

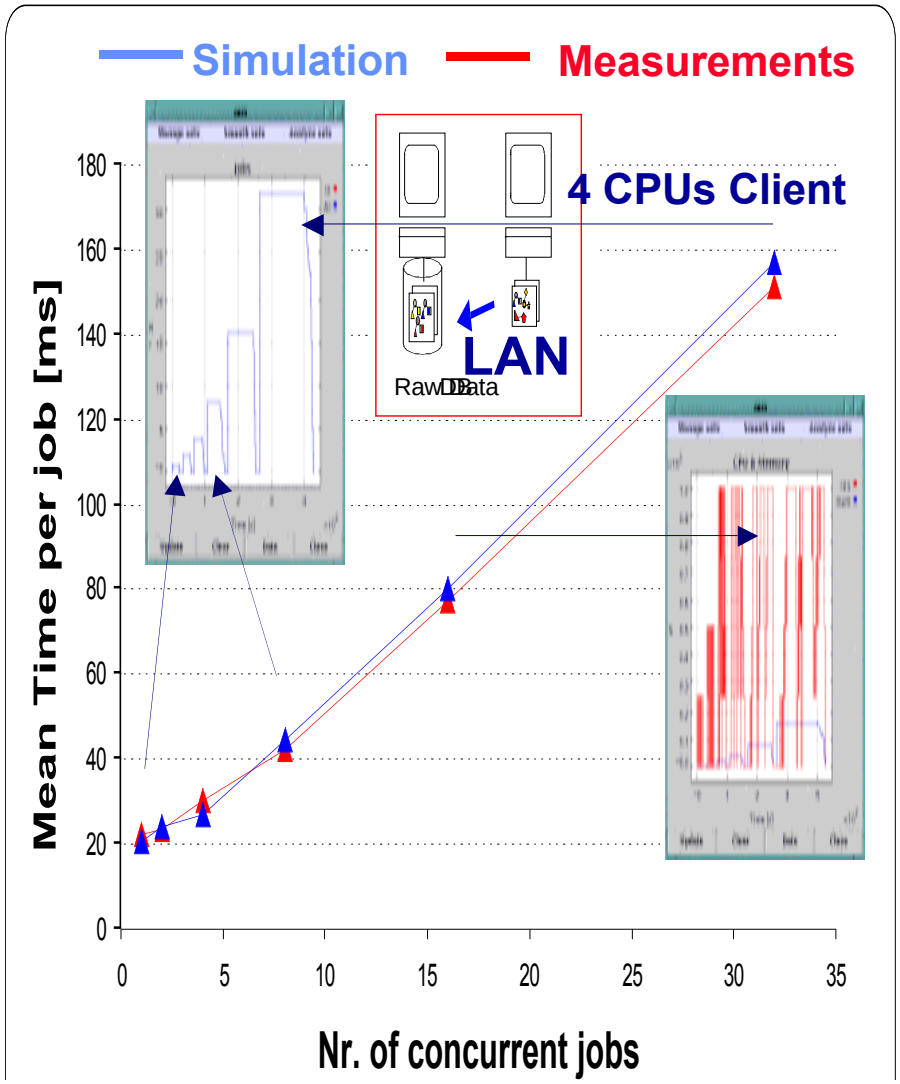
Simulation Validation: LAN Measurements (Y. Morita et al)

Machine A:
Sun Enterprise450 (400MHz 4x CPU)
Machine B:
Sun Ultra5 (270MHz): The Lock Server
Machine C:
Sun Enterprise 450 (300MHz 2x CPU)
Tests:
(1) Machine A local (2 CPU)
(2) Machine C local (4 CPU)
(3) Machine A (client) and Machine C (server)
number of client processes: 1, 2, 4, ..., 32

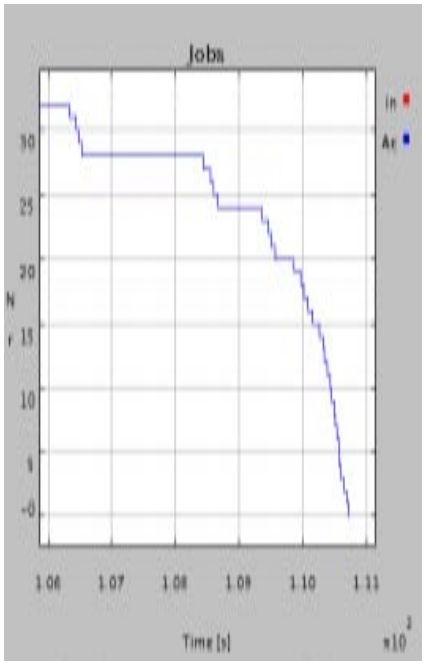


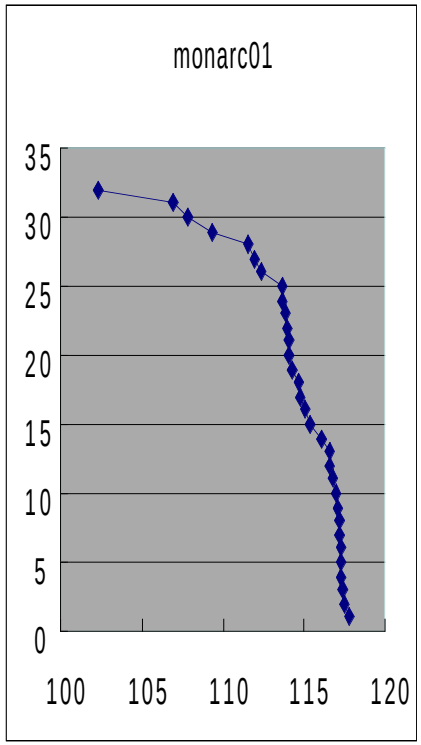


Validation Measurements
AMS Data Across a LAN



Distribution of 32 Jobs' Processing Time

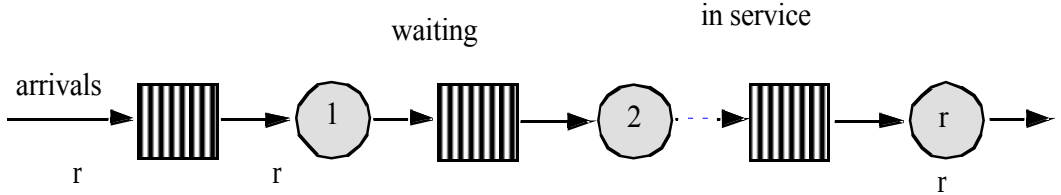




Simulation mean 109.5

Measurement mean 114.3

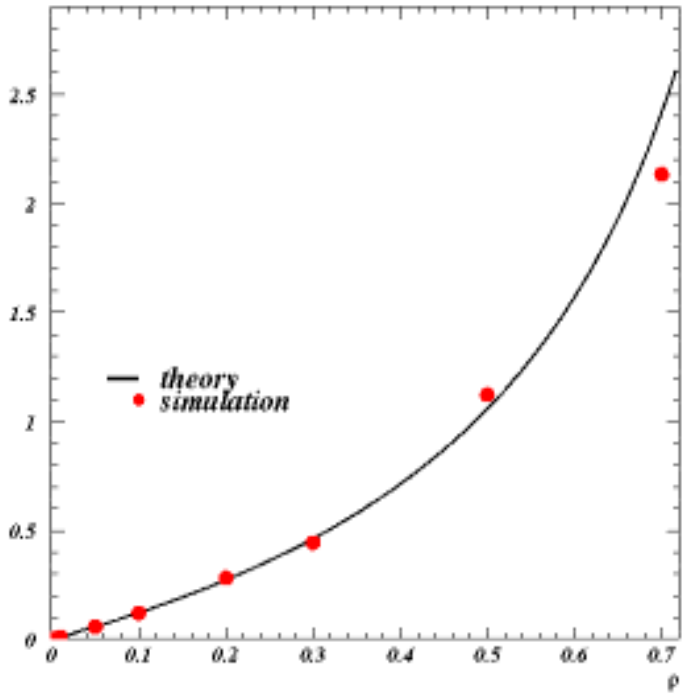
Queueing theory
M | M | 1 Network Queue Model



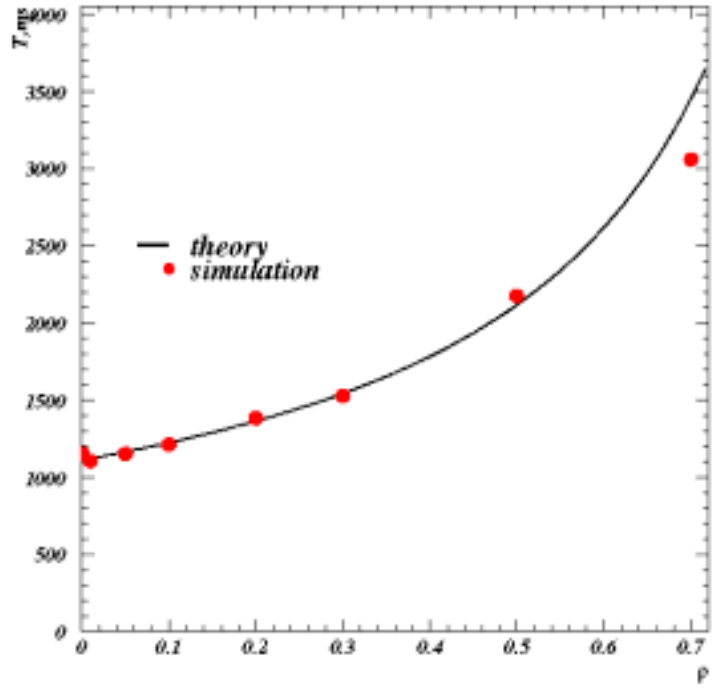
$$E[N] = \sum_{i=1}^{r} E[N_i] \qquad \sum_{i=1}^{r} \frac{\rho_i}{1-\rho_i} \quad \text{and}$$

$$E[R] = \sum_{i=1}^{r} E[R]_i \qquad \sum_{i=1}^{r} \frac{E[S_i]}{(1-\rho_i)}$$

Mean number of jobs vs utilisation



Mean response time vs utilisation



Modeling an AMS Across LANs and WANs

AMS page transfer latency is modeled into the simulation

Physical Bandwidth: B

Effective Bandwidth: B_{eff}

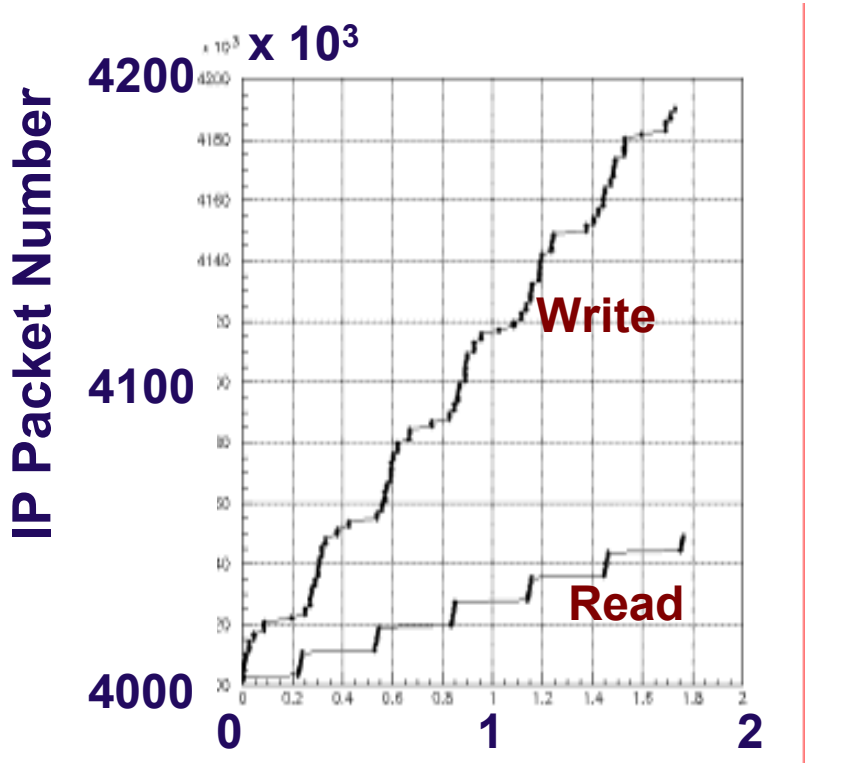
$\Delta T = \Delta t$ (*transfer*) + Δt (*handshake*)

= unit_size / B + RTT

$$\frac{B_{eff}}{B} = \frac{unit\_size}{unit\_size + B * RTT}$$

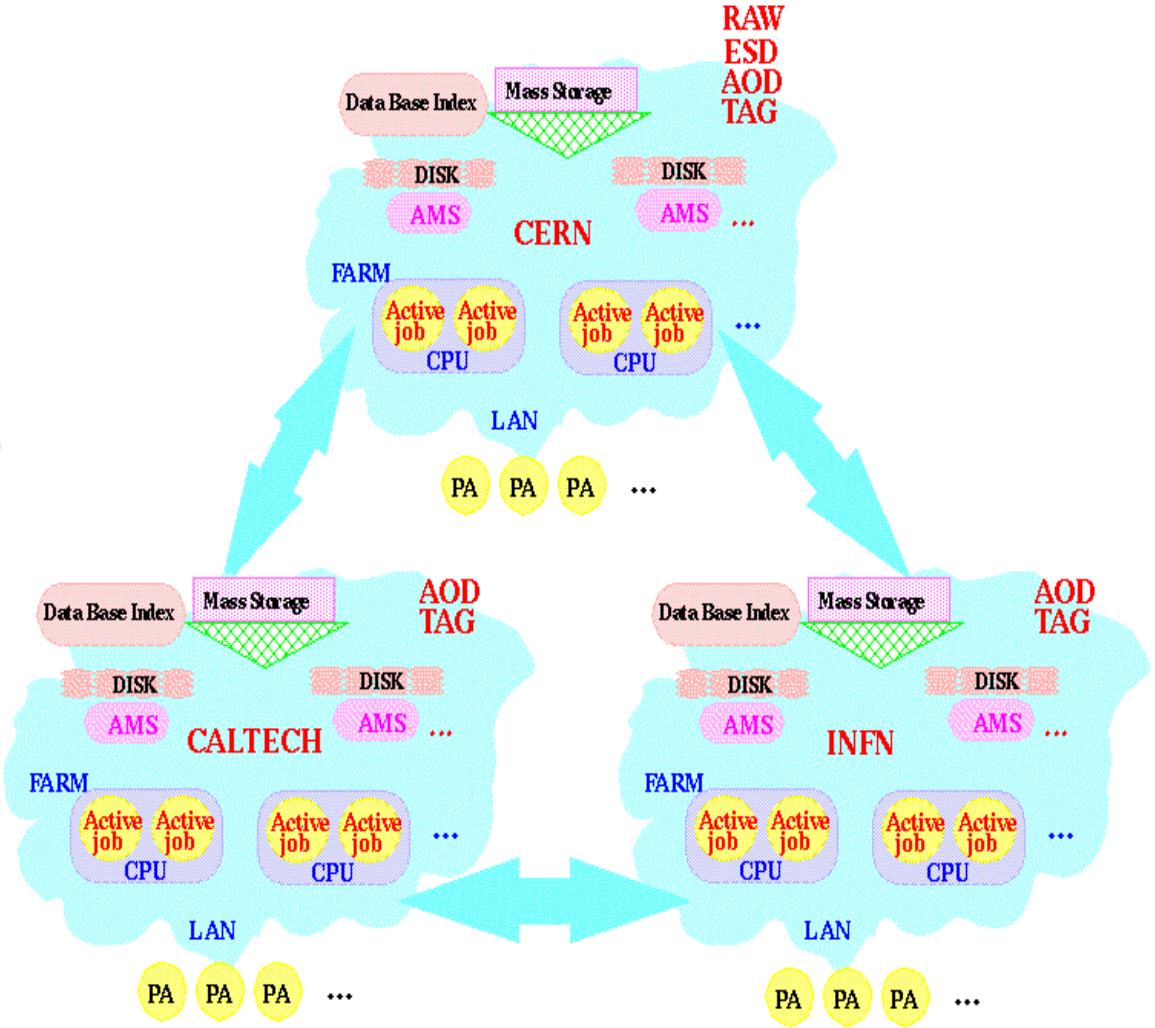
Study by H. Sato and Y. Morita
on CERN-KEK 2 Mbps Satellite Link
CHEP2000 Paper D235

AMS Packet Sequence

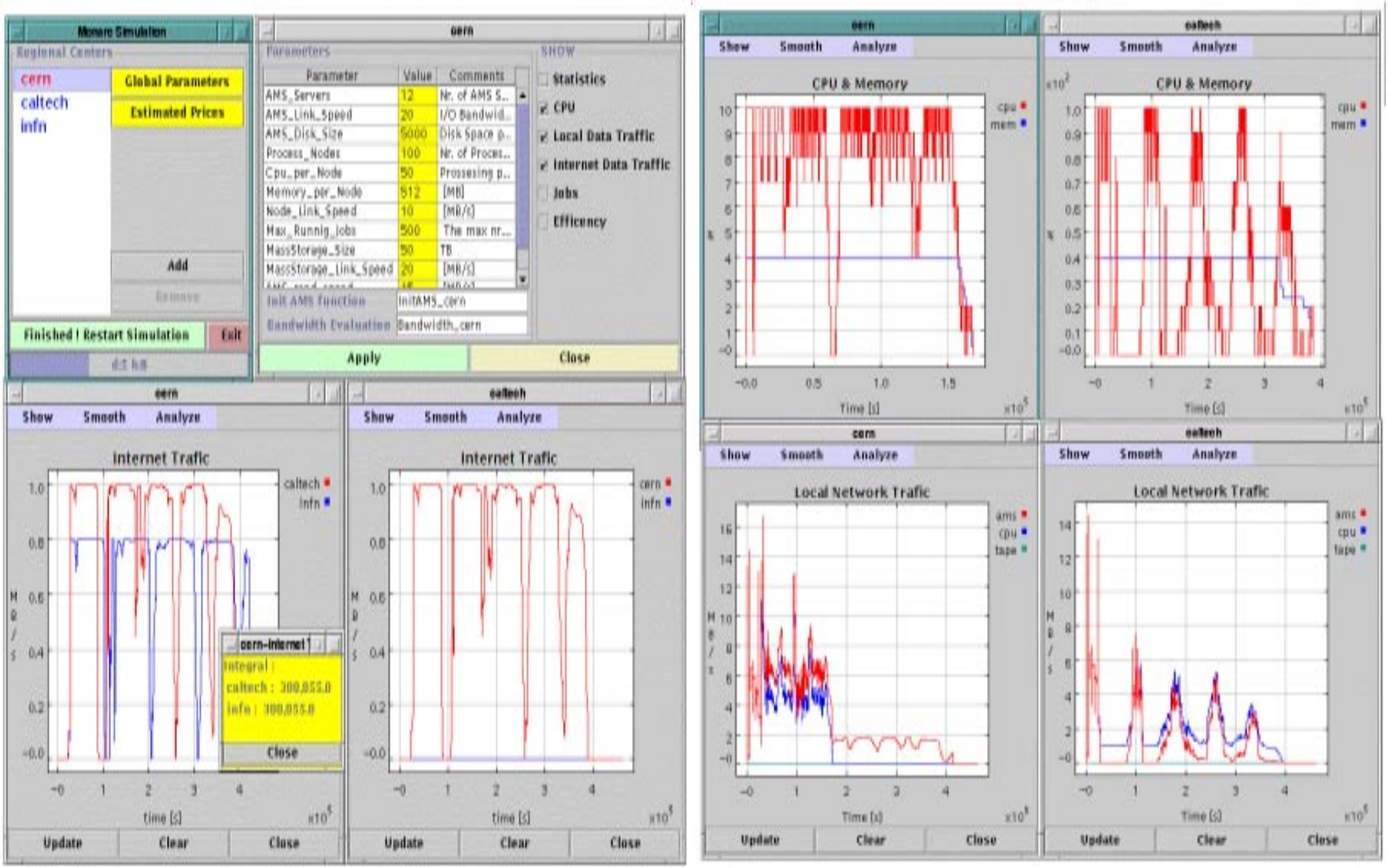


Example : Physics Analysis at Regional Centres

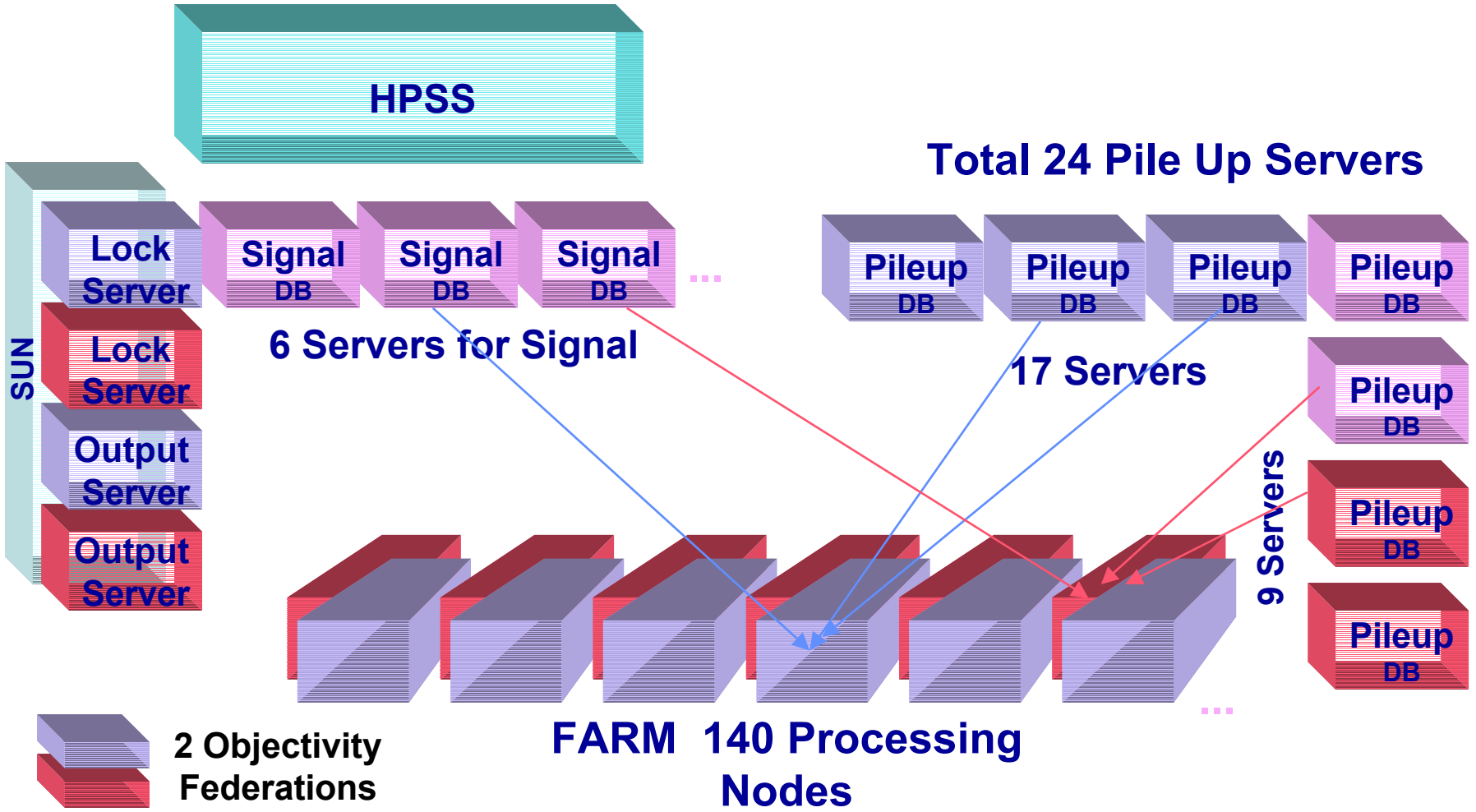
- Similar data processing jobs are performed in each of several RCs
- There is profile of jobs, each submitted to a job scheduler
- Each Centre has “TAG” and “AOD” databases replicated.
- Main Centre provides “ESD” and “RAW” data
- Each job processes AOD data, and also a a fraction of ESD and RAW data.



Example: Physics Analysis

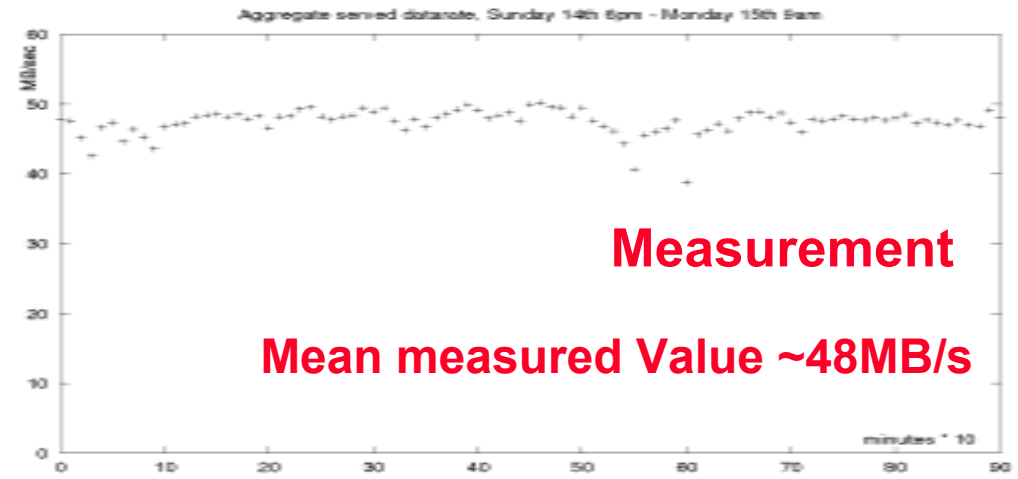


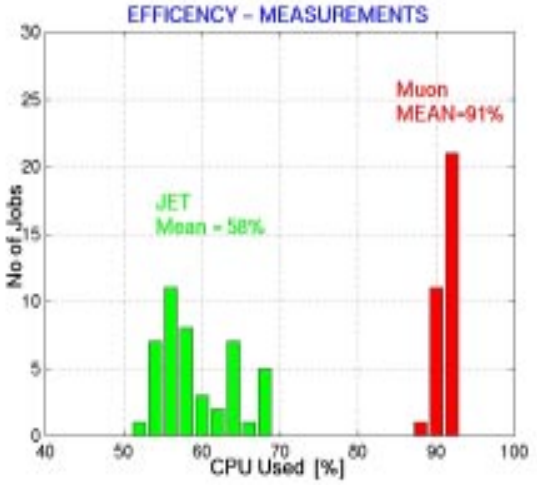
ORCA Production on CERN/IT-Loaned Event Filter Farm Test Facility

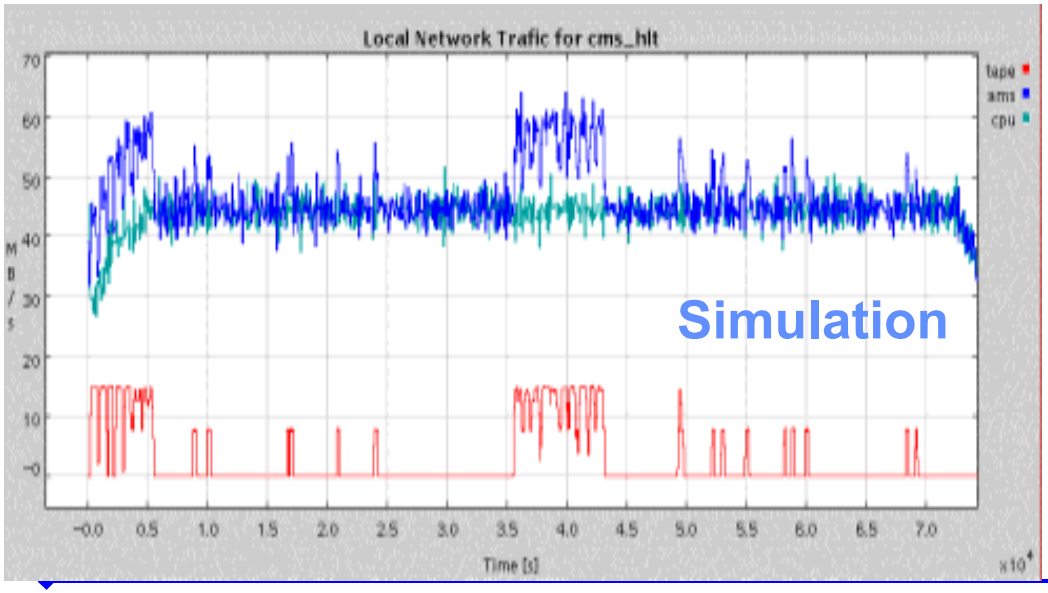


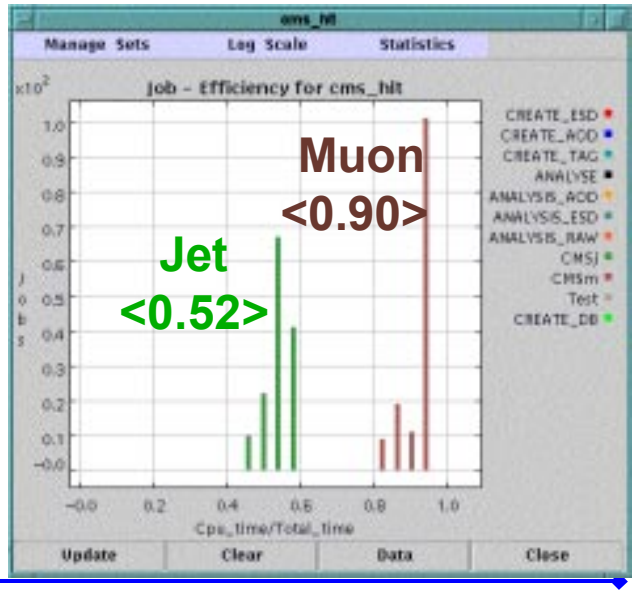
The strategy is to use many commodity PCs as Database Servers

Network Traffic & Job Efficiency

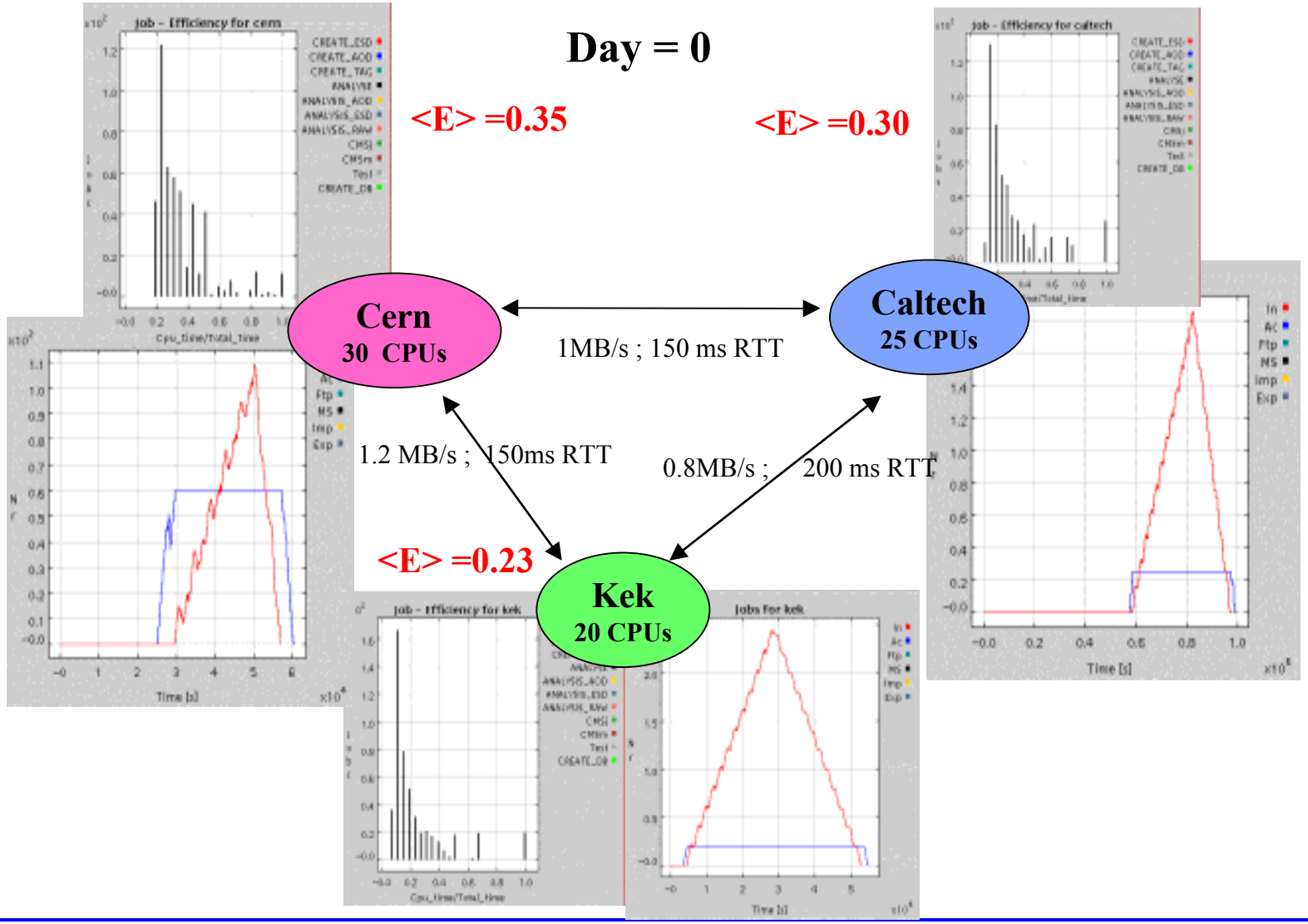




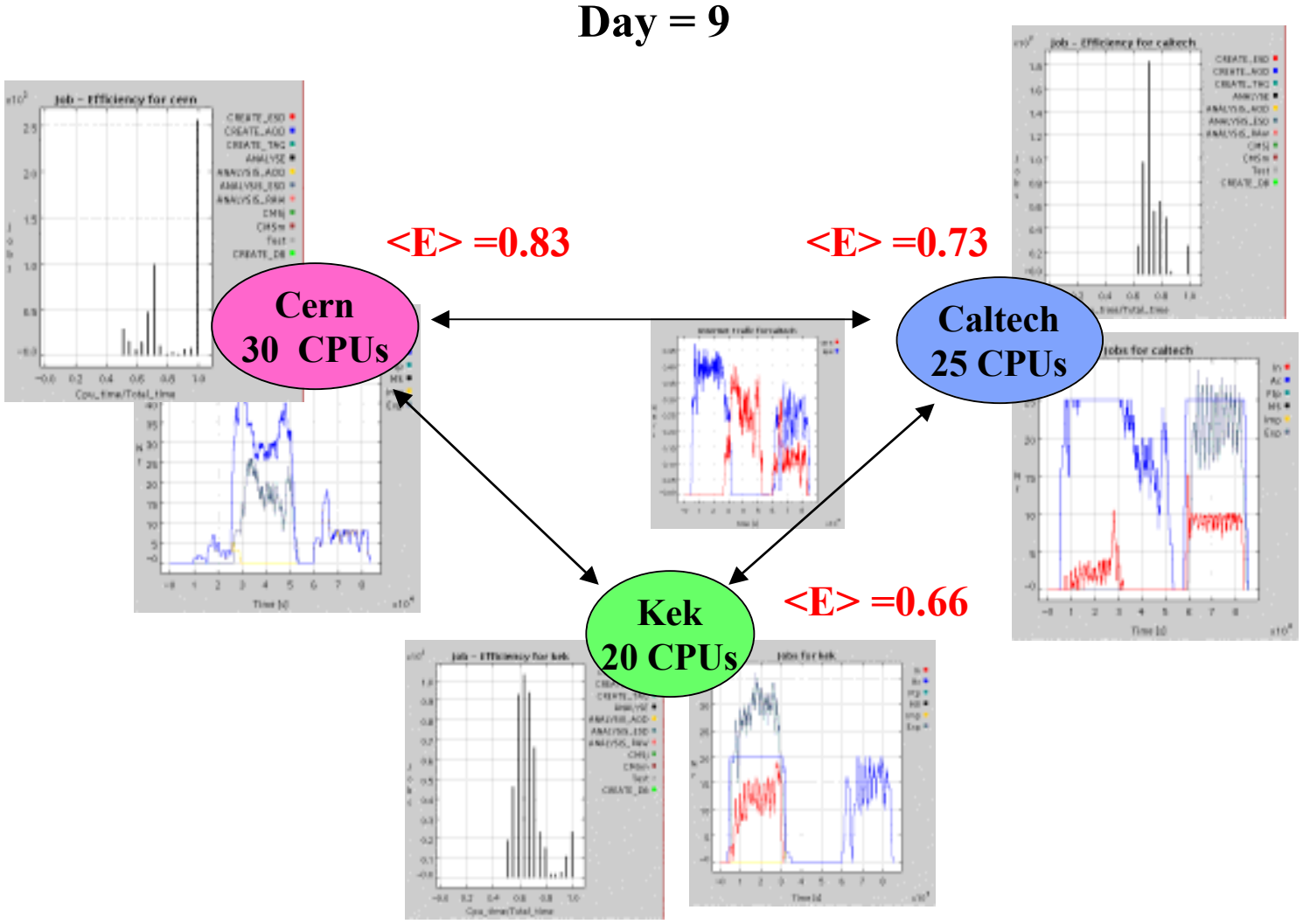




SONN: 3 RCs Learning to Export Jobs



SONN: 3 RCs Learning to Export Jobs




MONARC Simulation: I. LeGrand Workplan (2000)

May 2000: CMS HLT - simulation

- http://www.cern.ch/MONARC/sim_tool/Publish/CMS/publish/
- http://home.cern.ch/clegrand/MONARC/CMS_HLT/sim_cms_hlt.pdf

June 2000 Tape usage study

- http://www.cern.ch/MONARC/sim_tool/Publish/TAPE/publish/

Aug 2000 Update of the Simulation tool for large scale simulations.

- http://home.cern.ch/clegrand/MONARC/WSC/wsc_final.pdf
  (to be presented at the IEEE Winter Simulation Conference: WSC2000)
- http://home.cern.ch/clegrand/MONARC/ACAT/sim.ppt

Oct 2000 A study in using SONN for job scheduling

- http://www.cern.ch/MONARC/sim_tool/Publish/SONN/publish/
- http://home.cern.ch/clegrand/MONARC/ACAT/sonn.ppt

Nov 2000 Update of the CMS computing needs

- Based on the new requirements data, to update the baseline models for CMS computing

Dec 2000 Simulation of the current CMS Higher Level Trigger production

MONARC Simulation: I. LeGrand Workplan (2001)

Jan 2001 Update of the MONARC Simulation System

New release, including dynamic scheduling and replication modules (policies); Improved Documentation

Feb 2001 Role of Disk and Tapes in Tier1 and Tier2 Centers

More elaborate studies to describe Tier2-Tier1 interaction and to evaluate data storage needs

May 2001 Complex Tier0 - Tier1 - Tier2 simulation: Study the role of Tier2 centers

Aim is to perform a complete CMS data processing scenario including all major tasks distributed among regional centers

Jul 2001 Real SONN module for job scheduling; based on Mobile agents

Create a Mobile Agents framework able to provide the basic mechanism for scheduling between regional centers

Sep 2001 Add monitoring agents for network and system states based on (SNMP)

Collect system dependent parameters using SNMP and integrate them into the mobile agents used for scheduling

Dec 2001 Study of the correlation between data replication and job scheduling

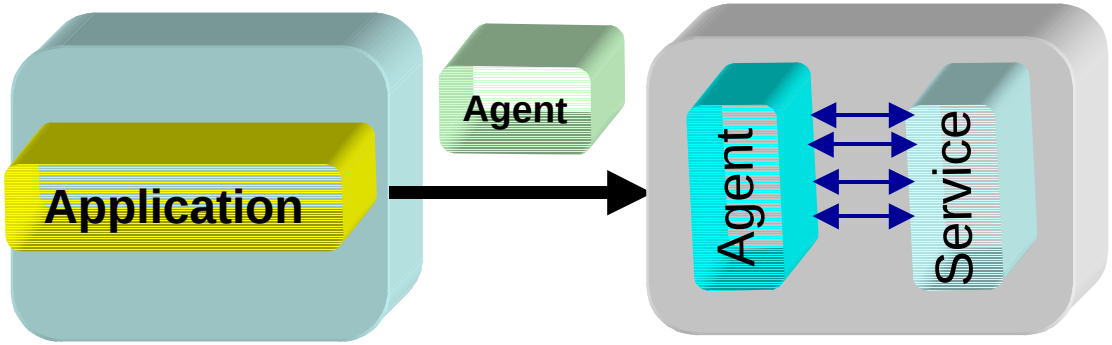
Combine the scheduling policies with data replication to optimize different cost functions; Integrate this into the Mobile Agents framework


MONARC Future: Some "Large" Grid Issues to Be Studied

- Query estimation and transaction-design for replica management
- Queueing and co-scheduling strategies
- Strategy for use of tapes
- Strategy for resource sharing among sites and activities
- Packaging of object-collections into blocks for transport across networks; integration with databases
- Effect on Networks of Large windows, QoS, etc.
- Behavior of the Grid Services to be developed

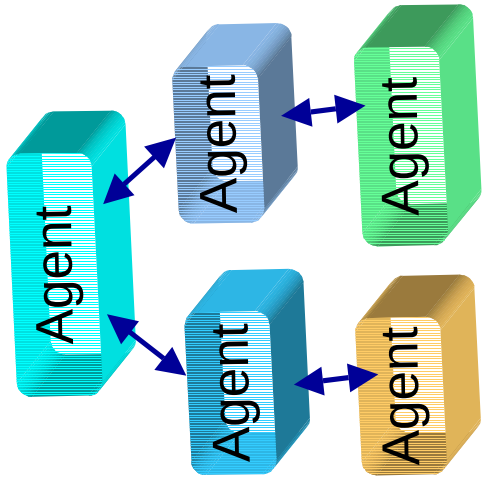
Beyond Traditional Architectures: Mobile Agents

"Agents are objects with rules and legs" -- D. Taylor



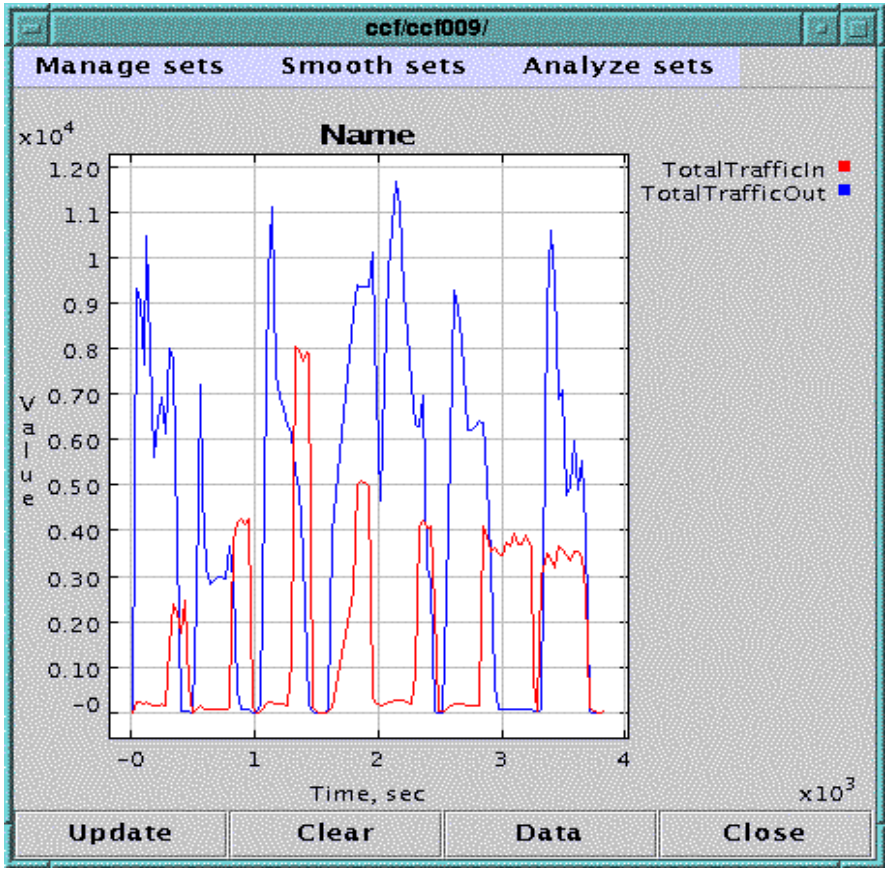
Mobile Agents: (Semi)-Autonomous, Goal Driven, Adaptive

- Execute Asynchronously
- Reduce Network Load: Local Conversations
- Overcome Network Latency; Some Outages
- Adaptive ➡ Robust, Fault Tolerant
- Naturally Heterogeneous
- Extensible Concept: Coordinated Agent Architectures

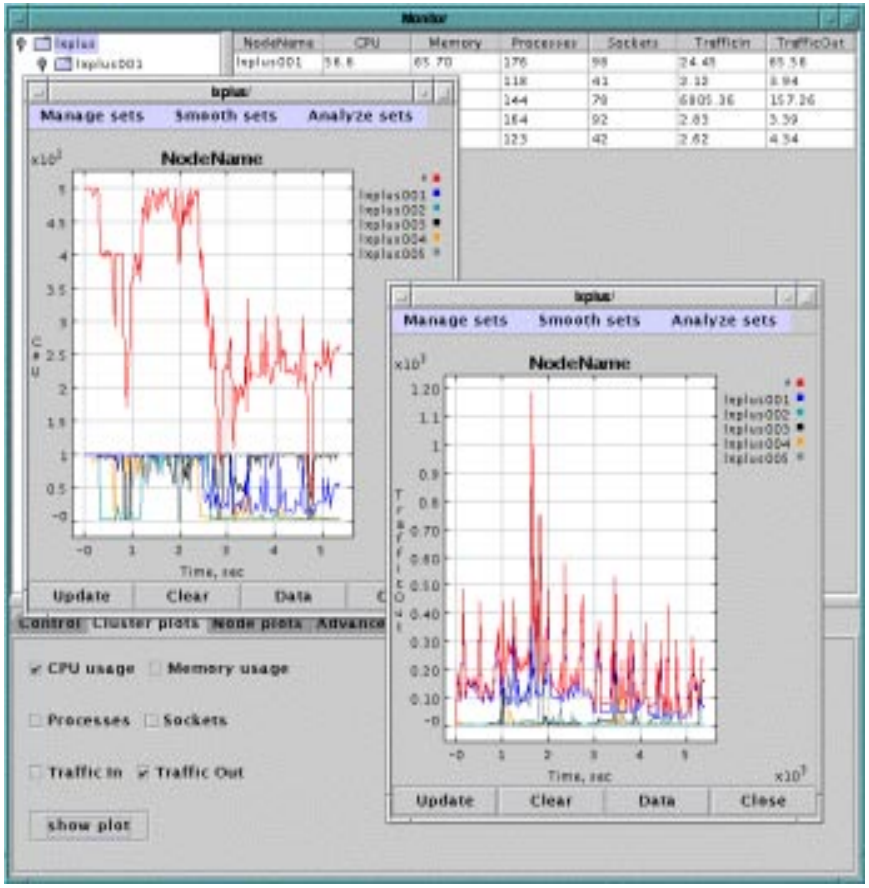


SNMP Based Monitoring Tool for Site and Network Activities

By I. Legrand, Caltech



Total IP traffic in the CERN domain



CPU usage & I/O per cluster

MONARC Status

- MONARC is on the way to specifying baseline Models representing cost-effective solutions to LHC Computing
- MONARC's Regional Centre hierarchy model has been accepted by all four LHC Experiments
  - ➡ And is the basis of HEP Data Grid work.
- A powerful simulation system has been developed, and is being used both for further Computing Model, Strategy Development, and Grid-component studies.
- There is strong synergy with other advanced R&D projects: PPDG, GriPhyN, EU HEP Data Grid, ALDAP and others.
- Example Computing Models have been provided, and are being updated
  - ➡ This is important input for the Hoffmann LHC Computing Review
- The MONARC Simulation System is now being applied to key Data Grid issues, and Grid-tool design and Development

MONARC: Current Status and How to Re-Start

- MONARC developed "static models" in which the resources were adjusted to meet the need
- The ORCA Spring 2000 production was the first large test that went beyond this
- It would be good to revive a MONARC Common Project, but with all the Grid Work Packages, I see little manpower outside of CMS available.
  - ➡ We have always had a "1+ FTE" effort on simulations
- We have to develop our own strategies
  - ➡ Adapted to CARF, and CARF-evolution
- We will be more able to move beyond the static strategies, with a (rather) CMS-specific effort
- Precondition: We need definite architectural proposals for how data should be structured and accessed
  - ➡ Else we cannot focus on a set of initial strategies, and alternatives, to be evaluated and improved.

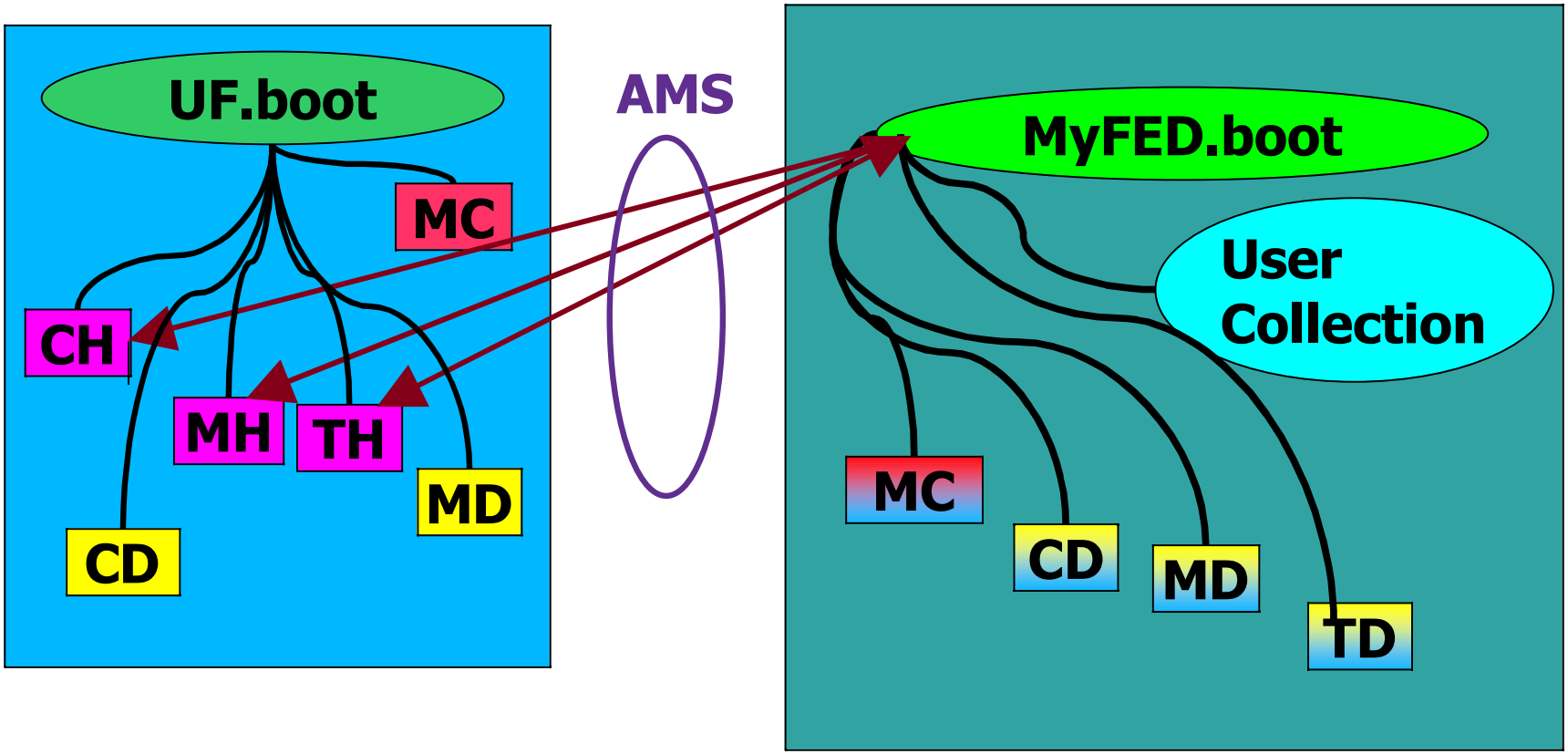
MONARC: How to Move Forward, and What is Needed (1)

- To get started we need an idea of what the persistent objects are in our REC, AOD, DPD, and TAG events;
  We need a schema, and we need to figure out:
  - If the schema leads to efficient access; else change it
  - When we should recluster; how often etc.
  - The time ("cost") for extracting and shipping out a collection
- Develop a complete concept of the workload presented by a user; and a complete concept of what he is trying to do
  - For example: Include his bringing some data to his desktop
- Develop (complete) the mix of users, their jobs and other tasks
- To do the above we need to know (roughly):
  - How the objects are stored at a site
  - Under which conditions they are accessed remotely; when they are shipped somewhere else for processing
  - Realistic target performance figures for discs when accessing data "randomly" or "serially"; and which are we doing under various conditions


MONARC: How to Move Forward (2)

- Once we understand how the data is structured, stored and accessed (as above), we then need to represent an interactive session:
  - What data is accessed and where; we need some policies and guidelines for this
  - Add in a load for “persistent” remote collaboration
- Once all of the above is done, with a good handle on the data flow and how the CPU is utilized we can then (and only then) move on to serious studies of workflow strategies and redirecting jobs
  - There is a DataGrid WorkPackage assigned to these issues, and we will have to coordinate with them on these aspects.

From UserFederation To Private Copy



ORCA 4 tutorial, part II - 14. October 2000